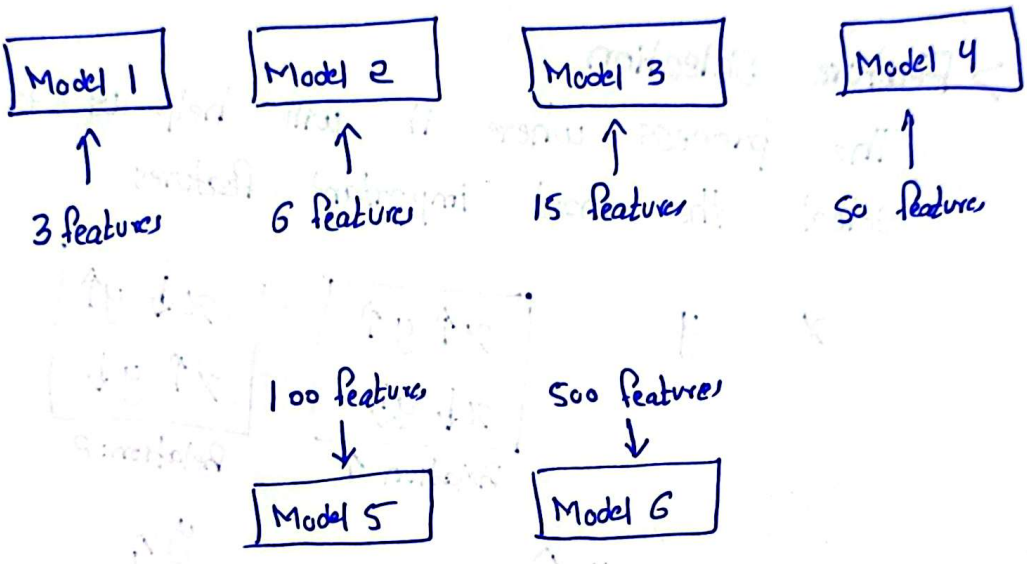
Principal Component Analysis

Also known as Dimensionality Reduction

→ Curse of Dimensionality

Dataset: 500 features



Model 1 : Accuracy

Model 2: Accuracy ↑↑

Model 3: Accuracy ↑↑↑

Model 4: Accuracy ↓↓

Model 5: Accuracy ↓↓↓

Model 6: Accuracy ↓↓↓↓

Two ways to remove curse of dimensionality

① Feature selection

② PCA => Feature Extraction

→ Feature Selection vs Feature Extraction

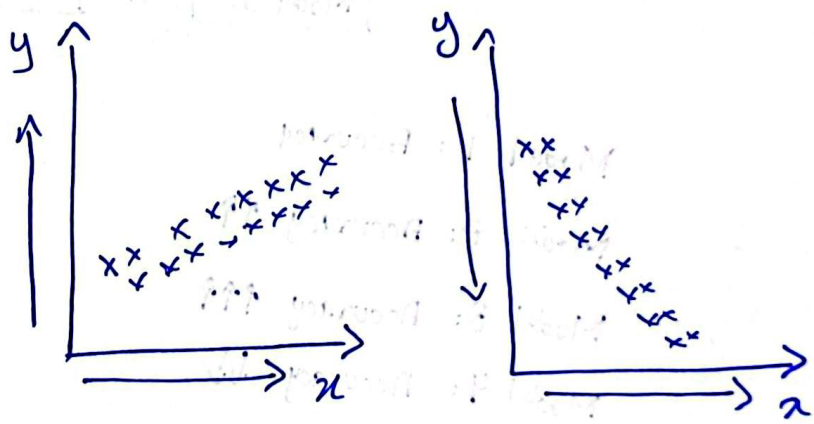
Why Dimensionality Reduction

1- Prevent Curse of Dimensionality
2- Improve performance of the model
3- Visualize the data

> Feature Selection

The process where it will help us to select the most important features

x	y	$x\uparrow\ y\uparrow$ $x\downarrow\ y\downarrow$	$x\downarrow\ y\uparrow$ $x\uparrow\ y\downarrow$
		Relation 1	Relation: 2



$$cov(x,y) = \frac{\sum_{i=1}^{n} (x-\bar{x})(y-\bar{y})}{N-1}$$

$cov(x,y) = +ve \Rightarrow$ Relation: 1

$cov(x,y) = -ve \Rightarrow$ Relation: 2

$cov(x,y) \approx 0 \Rightarrow$ No relation

Correlation

$$\text{Pearson Correlation} = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$

$$\text{If} \quad = -1 \text{ to } 1$$

More towards +1, more positive correlation

More toward -1, more negative correlation

> Feature Extraction

Room Size No.of rooms Price

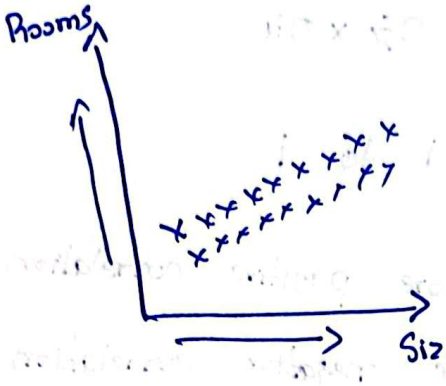
Transformation to extract new feature

Room Size
No of room
} Transformation => House Size

House Size Price

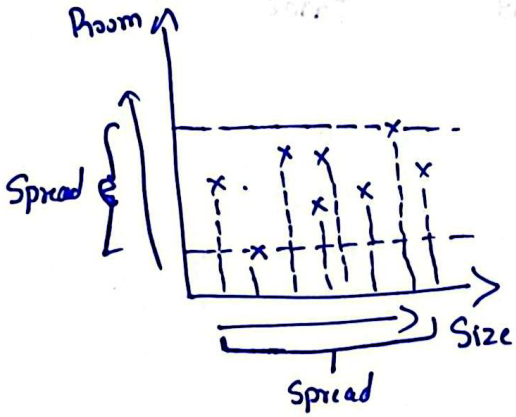
⟶ Geometric Intuition

Size of house | no.of rooms | Price



2 dimensions ⟶ 1 dimension

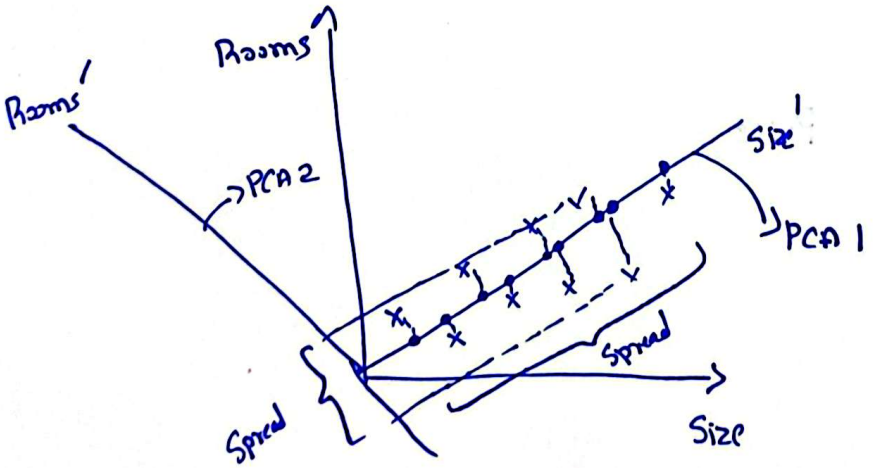
Projecting Data Points



2D ⟶ 1D

* Loss of information (no. of rooms)

PCA



Eigen decomposition on some matrix

⇑ Transformation

maximum variance is captured

var(PCA1) > var(PCA2) Much information is not lost

2D ⟶ 1D